# Getting Started with ESS-DIVE's Reporting Formats

Emily Robles, Joan Damerow

ESS-DIVE BERKELEY LAB NERSC NCEAS DataONE U.S. DEPARTMENT OF ENERGY Office of Science

Cyberinfrastructure Working Group Meeting 2024

# Presenters

**Emily Robles**
Senior Research Associate

**Joan Damerow**
Community Engagement Lead, Research Scientist

# Overview

## 1. File-level Metadata (FLMD)

- Overview, requirements, use
- Hands-on practice

## 2. CSV Reporting Format

- Overview, requirements, use
- Hands-on practice

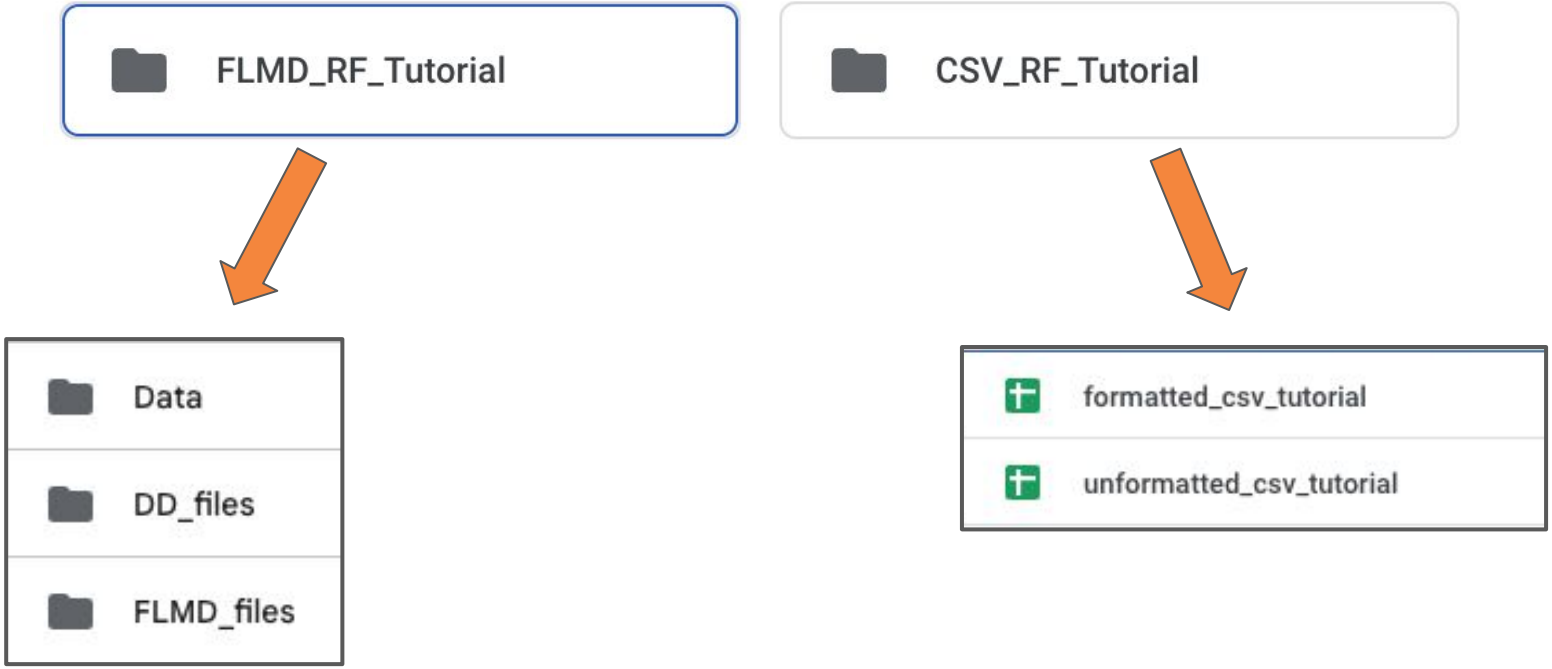
## 3. Samples Reporting Format

## 4. Publishing Datasets with Reporting Formats

- Review and publication workflow
- Common errors and potential for tools

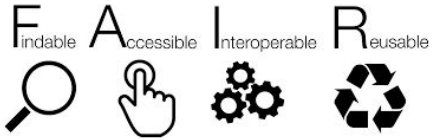
***Questions and discussion encouraged!***

# Access to Hands-on Content



# Benefits of Reporting Formats

The standardized data and metadata templates improve both **human-** and **machine-readability**




- Makes data more Findable, Accessible, Interoperable, Reusable
- Set consistent methods of reporting data within a project
- Allows scientists to **easily work across multiple datasets**
- Planned ESS-DIVE tools for advanced data search, integration, and visualization **will leverage reporting formats**

# ESS-DIVE Reporting Formats

**Data Files**

CSV Guidelines

Velliquette, Heinz, Devarakonda (ORNL)

Model Data Archiving

Simmonds (LBNL)

Soil Respiration

Bond-Lamberty, Pennington (PNNL)

Leaf Physiology

Rogers, Ely (BNL)

Hydrologic Monitoring

Goldman (PNNL)

Water/Soil Chemistry

Boye (SLAC)

Amplicon Sequencing

Weisenhorn (ANL)

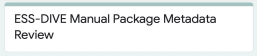
UAS

Serbin, Ely (BNL)

**Metadata**


Package Level Metadata/JSON-LD

Agarwal, Hendrix (LBNL)




Package Metadata Quality

Damerow (LBNL)

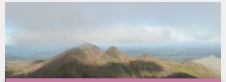
File-level Metadata

Velliquette, Heinz, Devarakonda (ORNL)

SESAR
SYSTEM FOR EARTH SAMPLE REGISTRATION

Sample IDs/Metadata

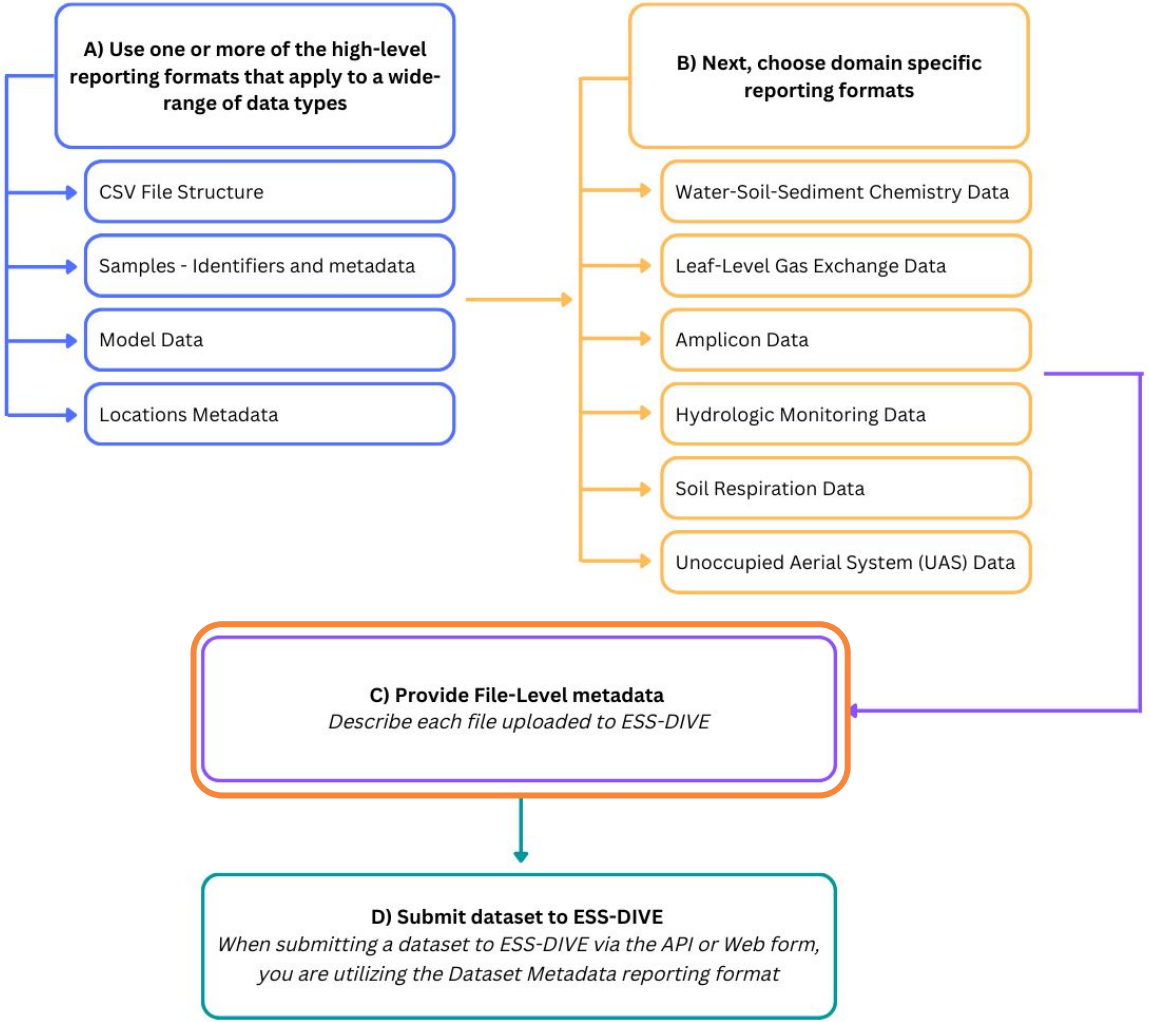
Damerow (LBNL)



Locations metadata

Crystal-Ornelas (LBNL)

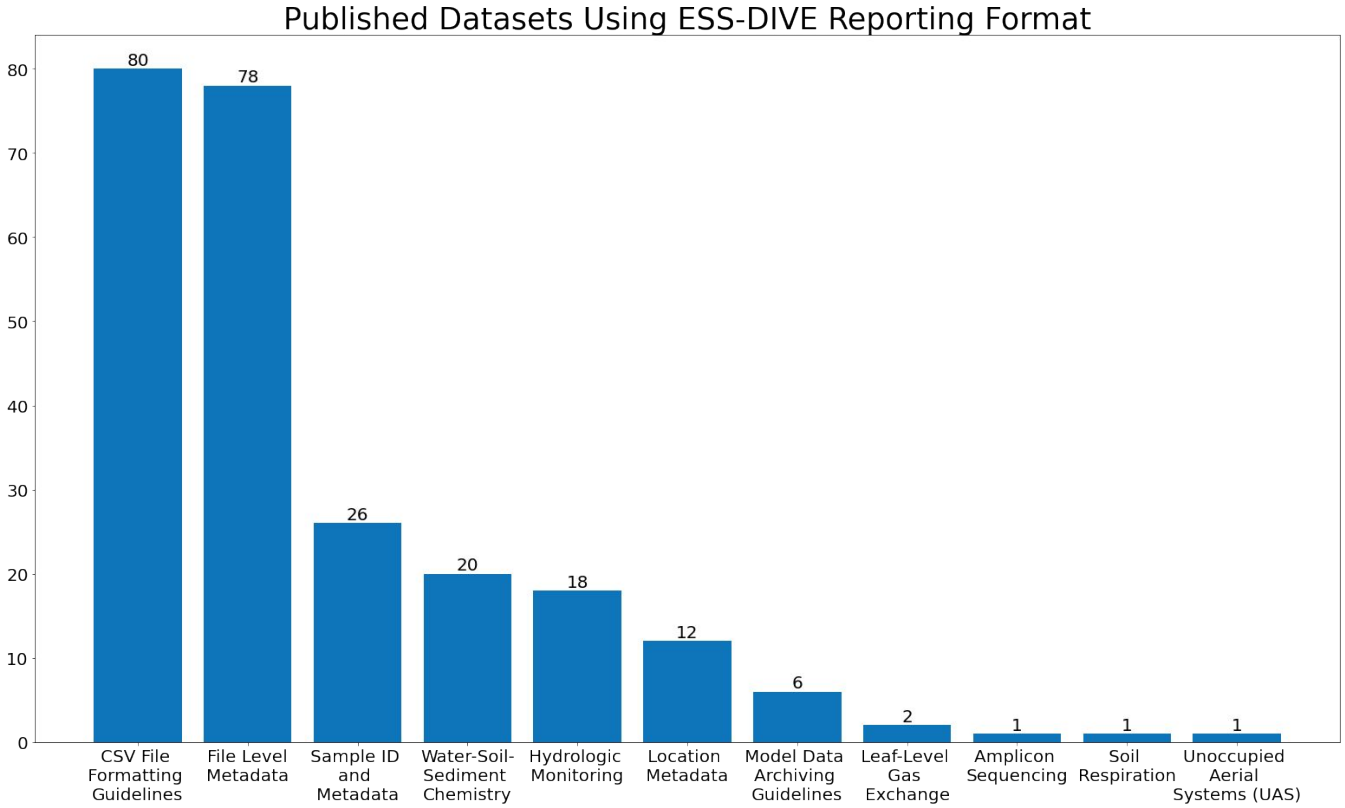
Crystal-Ornelas, R. et al. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. Sci Data 9, 700 (2022). https://doi.org/10.1038/s41597-022-01606-w




***Reporting formats offer multiple levels of formatting guidance***

# Status of Reporting Format Datasets

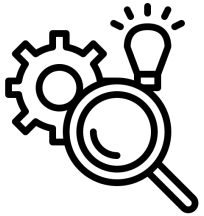
Early adoption has been key to the development of tools and features

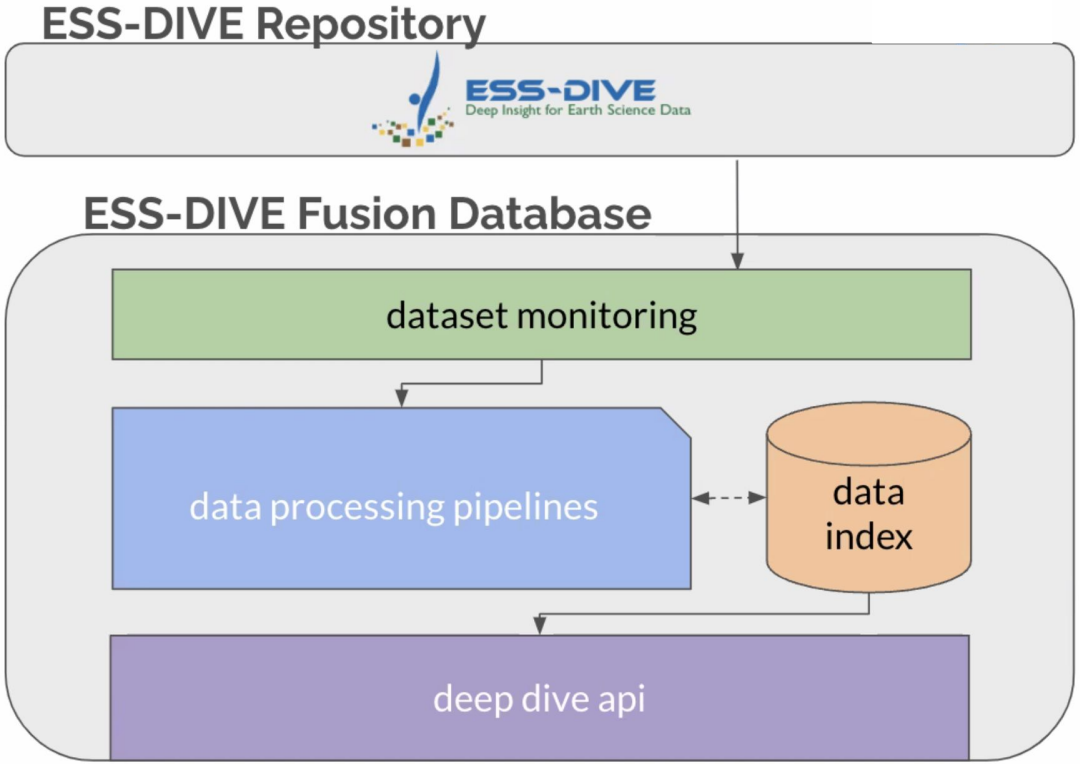



***96 datasets using reporting formats are publicly available***

# Published Datasets Enable Enhanced Search

Completed FLMD and data dictionary files enable Fusion DB to **find** and **extract** your data from parsable CSV files in published datasets





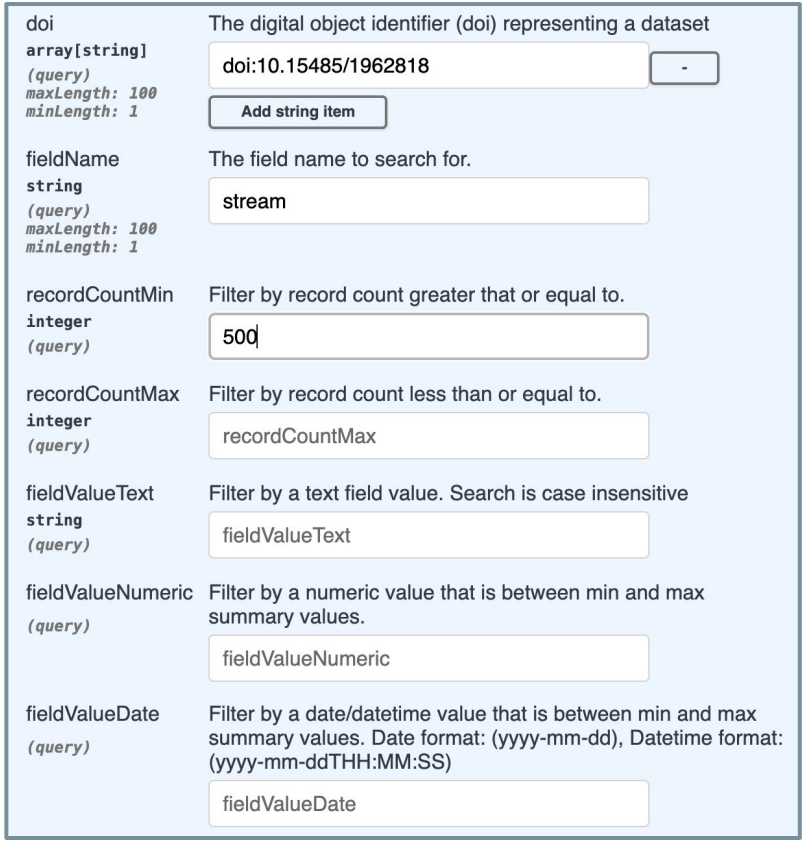
# Published Datasets Enable Enhanced Search

Use of reporting formats enables enhanced search through the **Deep Dive API**

- Separate from ESS-DIVE main search
- Searches data **within dataset files**

Published datasets that employ reporting formats are instrumental to enabling advanced search

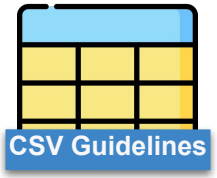
***Interactive API at [fusion.ess-dive.lbl.gov](fusion.ess-dive.lbl.gov)***



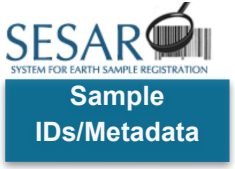
# Getting Started

***Hands-on practice with broadly applicable reporting formats***


Metadata at the file-level for all datasets and data types




Guidelines for all data in CSV tabular format




Guidelines for archival of sample data with IGSN IDs


***Later today!***

Model code management and model data archiving

# File-level Metadata Reporting Format

Purpose, requirements, and hands-on exercise

# The Role of File-level Metadata

File-level metadata provides the information needed to understand, parse, and extract data from files. It consists of two files:

1. File-level Metadata File (FLMD)
   - Each row contains information about a **file** within the dataset

2. Data Dictionary File (dd)
   - Each row contains the **header row/column information** of **individual files**

***The ESS-DIVE Fusion DB uses the FLMD reporting format to parse CSV files and enable advanced search capabilities***

# FLMD v2.0 Requirements and Recommendations

## Required

**File Name**

- Name of associated file

**File Description**

- Brief description of file that distinguishes it from other files
- Information about data type

## Recommended

**Standard**

- State if any data or metadata standard was applied to the data file
- If reporting formats were used, use defined terms provided

## Optional

- **Header rows**
- **Column or row name position**
- File Version
- Data Orientation
- Notes

# NEW FLMD Header Rows and Position

Optional FLMD fields allow for handling of additional header rows/columns before ***and*** after the column or row names

| | | | | | |
|---|---|---|---|---|---|
| Information before headers | | | | | |
| Additional information | | | | | |
| One more row of information | | | | | |
| **leaf** | **date** | **measurement_time** | **conductance** | **temperature** | **leaf_sensor** |
| 1 | 2017-06-23 | 7.0 | 318.8 | 37.3 | 48.5 |
| 2 | 2017-06-23 | 7.0 | 277.1 | 35.9 | 62.8 |
| 3 | 2017-06-23 | 7.0 | 267.3 | 36.1 | 62.9 |
| 4 | 2017-06-23 | 7.0 | 200.5 | 36.2 | 68.0 |

| **Sample_Name** | **Date** | **Time_Collected** | **Water_Temperature** | **Dissolved_Oxygen** |
|---|---|---|---|---|
| N/A | YYYY-MM-DD | hh:mm | Degrees_Celsius | miligrams_per_Liter |
| Site 1 Sample | 2022-01-12 | 13:05 | 22 | 10.05 |
| Site 2 Sample | 2022-01-12 | 13:50 | 20.7 | -9999 |
| Site 3 Sample | 2022-01-12 | 14:22 | 19.7 | -9999 |
| Site 4 Sample | 2022-01-12 | 14:56 | -9999 | 10.56 |

# Header Rows and Position

**header_rows**

- Used when rows after variable/header names and before data
- Provide the number of header rows that occur after the column or row names in a file and before the data begins

| Sample_Name | Date | Time_Collected | Water_Temperature |
|---|---|---|---|
| N/A | YYYY-MM-DD | hh:mm | Degrees_Celsius |
| Site 1 Sample | 2022-01-12 | 13:05 | 22 |
| Site 2 Sample | 2022-01-12 | 13:50 | 20.7 |
| Site 3 Sample | 2022-01-12 | 14:22 | 19.7 |
| Site 4 Sample | 2022-01-12 | 14:56 | -9999 |

***header_rows = 1***

# Header Rows and Position

**column_or_row_name_position**

- Used when rows before variable/header names
- Provide the row or column number that contains the header names
- If not included, it will be assumed that header names are in row 1 (for horizontal orientation)

| | | | | |
|---|---|---|---|---|
| Location information: Site 1 had pH monitor issues for sample on 2022-01-18 15:32 | | | | |
| Contact emilyarobles@lbl.gov for more information about dataset | | | | |
| Sample_Name | DateTime_Start | DateTime_End | Location_ID | Water_Temp |
| Site_1_2022-01-18 | 2022-01-18 15:32 | 2022-01-18 16: | Site_1 | 10.2 |
| Site_3_2022-01-18 | 2022-01-18 15:23 | 2022-01-18 15: | Site_3 | 10.4 |
| Site_7_2022-01-18 | 2022-01-18 15:11 | 2022-01-18 15: | Site_7 | 10.1 |
| Site_8_2022-01-18 | 2022-01-18 15:18 | 2022-01-18 15: | Site_8 | 10.7 |

***column_or_row_name_position = 3***

***Important: If row is commented out (preceded by hash mark) the row should not be counted***

# NEW Controlled Standard Names

**standard**

- Note if any standard formats are being used in a data file
- Now have a list of names to use for the ESS-DIVE reporting formats

| | | | | |
|---|---|---|---|---|
| | | ESS-DIVE FLMD v1 | | |
| ESS-DIVE Model Data v1 | ESS-DIVE Soil Respiration v1 | ESS-DIVE Water-Soil-Sediment Chem v1 | ESS-DIVE Leaf-level Gas Exchange v1 | ESS-DIVE UAS v1 |
| ESS-DIVE Amplicon v1 | ESS-DIVE Hydrologic Monitoring v1 | ESS-DIVE Sample v1 | ESS-DIVE CSV v1 | ESS-DIVE Location v1 |

***Reminder: Include the name(s) of reporting format(s) used in the package-level keywords when submitting your dataset***

# Data Orientation

Two options for noting data orientation for CSV files within the file-level metadata:

1. Horizontally with Names at the top of each column

OR

2. Vertically with Names at the start of each row.

***How are data being added to my data table?***

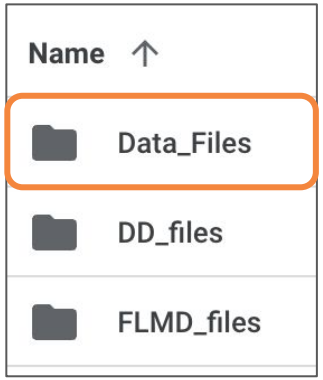
**New row = Horizontal orientation**

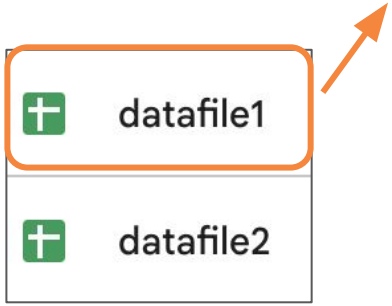
| area | plot_type | Latitude | Longitude | year | CH4_flux |
|---|---|---|---|---|---|
| Site 6 | CLC1 | 71.29573 | -156.66473 | 2010-07-07 | 91.8 |
| Site 6 | CLC2 | 71.29571 | -156.66469 | 2010-07-07 | 54.3 |

**New column = Vertical Orientation**

| area | Site 6 | Site 6 | Site 6 |
|---|---|---|---|
| plot_type | CLC1 | CLC2 | CLC3 |
| Latitude | 71.29573 | 71.29571 | 71.2957 |
| Longitude | -156.66473 | -156.66469 | -156.66467 |
| year | 2010-07-07 | 2010-07-07 | 2010-07-07 |
| CH4_flux | 91.8 | 54.3 | 63.9 |

# Practice: Filling out file-level metadata

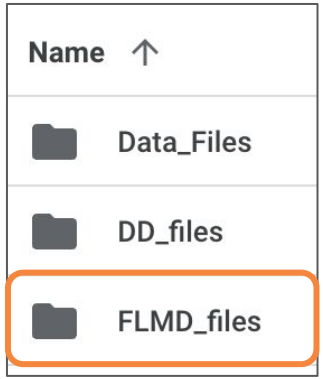


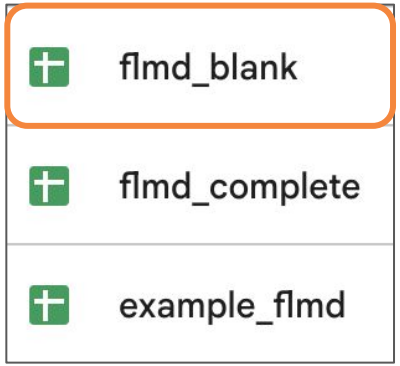




Example data file 1

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Sample_Name | DateTime_Start | DateTime_End | Location_ID | Latitude | Longitude | Water_Temper |
| 2 | Site_1_2022-01-18 | 2022-01-18 15:32 | 2022-01-18 16:03 | Site_1 | 38.14637 | -121.25532 | 10.2 |
| 3 | Site_3_2022-01-18 | 2022-01-18 15:23 | 2022-01-18 15:45 | Site_3 | 38.14824 | -121.26637 | 10.4 |
| 4 | Site_7_2022-01-18 | 2022-01-18 15:11 | 2022-01-18 15:32 | Site_7 | 38.1497 | -121.29353 | 10.1 |
| 5 | Site_8_2022-01-18 | 2022-01-18 15:18 | 2022-01-18 15:44 | Site_8 | 38.14943 | -121.29981 | 10.7 |
| 6 | Site_9_2022-01-18 | 2022-01-18 15:23 | 2022-01-18 15:55 | Site_9 | 38.1533 | -121.3 | 10 |

# Practice: Filling out file-level metadata




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | file_name | file_description | standard | file_version | data_orientation | notes |
| 2 | | | | | | |
| 3 | | | | | | |

# Practice: Filling out file-level metadata

**Data File 1 metadata tab**

| | A | B |
|---|---|---|
| 1 | File Version | 1 |
| 2 | File Description | Data file containing water quality measurements |
| 3 | Standard | ESS-DIVE CSV v1 |

**File-level Metadata Template (flmd_blank)**

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | file_name | file_description | standard | file_version | data_orientation | notes |
| 2 | | | | | | |
| 3 | | | | | | |

# Practice: Filling out file-level metadata

## Required

**File Name**

- Name of associated file

**File Description**

- Brief description of file that distinguishes it from other files
- Information about data type

## Recommended

**Standard**

- State if any data or metadata standard was applied to the data file (including reporting formats)

## Optional

- Header rows
- Column or row name position
- File Version
- Data Orientation
- Notes

*Complete for both data files*

# Completed FLMD

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | file_name | file_description | standard | file_version | data_orientation | notes |
| 2 | datafile1.csv | Data file containing water quality measurements. Additionally, metadata is contained related to the data and the file. | ESS-DIVE CSV v1 | 1 | horizontal | Data processing details in methods.pdf |
| 3 | datafile2.csv | Data file containing water quality measurements. Additionally, metadata is contained related to the data and the file. | ESS-DIVE CSV v1 | 1 | horizontal | Data processing details in methods.pdf |

# Data dictionary fields

Used to describe each field in CSV/tabular data files

## Required

**Column or Row Name**

- Each column or row name from the data file

**Unit**

- Provide variable units of measurement or "N/A" if units aren't applicable

**Definition**

- A complete *unambiguous* description of column or row

## Optional

**Column or Row Long Name**

- Longer human-readable column or row name

**Data Type**

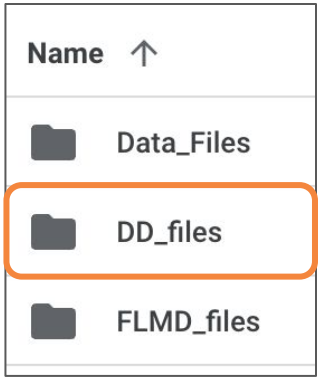
- Define the data type for each column (e.g. text, numeric, date)

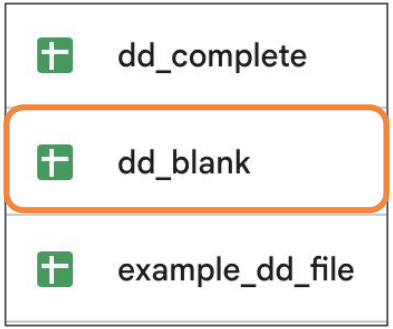
**Missing Value Code**

- Define the missing value codes used for a specific field

*Included as a CSV file **in addition** to your FLMD file and other reporting formats if applicable.*

# Practice: Complete a data dictionary




| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | column_or_row_name | unit | definition | data_type | missing_value_codes |
| 2 | | | | | |
| 3 | | | | | |

# Data Dictionary

## Data File

| Sample_Name | DateTime_Start | DateTime_End | Location_ID | Latitude | Longitude | Water_Temperature | pH | Dissolved_Oxygen | Turbidity | Notes |
|---|---|---|---|---|---|---|---|---|---|---|
| Site_1_2022-01-18 | 2022-01-18 15:32 | 2022-01-18 16:03 | Site_1 | 38.14637 | -121.25532 | 10.2 | -9999 | 11.5 | 1.6 | pH meter did r |
| Site_3_2022-01-18 | 2022-01-18 15:23 | 2022-01-18 15:45 | Site_3 | 38.14824 | -121.26637 | 10.4 | 8.4 | 11.3 | 2.1 | N/A |
| Site_7_2022-01-18 | 2022-01-18 15:11 | 2022-01-18 15:32 | Site_7 | 38.1497 | -121.29353 | 10.1 | 8.3 | 10.9 | 1.9 | N/A |

## Data Dictionary Template (dd_blank)

For each column* in your data files, you should have a row in your data dictionary

| A | B |
|---|---|
| column_or_row_name | unit |

Repeated variables only need to be entered once

* or row if vertical orientation

# Data Dictionary

## Completed Data Dictionary

| A | B | C | D |
|---|---|---|---|
| column_or_row_name | unit | definition | data_type |
| Sample_Name | text | name of sample | text |
| DateTime_Start | YYYY-MM-DD hh:mm | time at start of monitoring | N/A |
| DateTime_End | YYYY-MM-DD hh:mm | time at completion of monitoring | N/A |
| Location_ID | N/A | location name | text |
| Water_temperature | Degrees_Celsius | temperature of water sample collected | numeric |
| pH | N/A | measured pH of the watere sample | numeric |
| Dissolved_Oxygen | milligrams_per_liter | measured Dissolved Oxygen of the water sample | numeric |
| Turbidity | NTU | measured turbidity of the water sample | numeric |
| nitrates | milligrams_per_liter | measured nitrates of the water sample | numeric |
| Latitude | decimal degrees | latitude of the location | numeric |
| Longitude | decimal degrees | longitude of the location | numeric |
| Notes | text | notes associated with a sample or measurement | text |

# FLMD documentation and instructions

Guide to Using ESS-DIVE

Main Website | Search Data | Submit Data | Contact Us

Search...

Welcome

Frequently Asked Questions

SUBMIT DATA

Get Started

Register to Submit Data

Dataset Requirements

Describe & Format Datasets

Submit Data with Online

## File Level Metadata

| File Level Metadata | Status: **READY TO USE** |
|---|---|
| Purpose | Provide metadata for each data file uploaded to ESS-DIVE |
| Who should use | Anyone submitting files to ESS-DIVE, regardless of file type |
| Documentation | GitHub or GitBook |

Copy link

ON THIS PAGE

File Level Metadata

CSV File Structure

Sample ID Metadata

Soil Respiration

Leaf-level Gas Exchange

Hydrologic Monitoring

File-level metadata reporting format

Overview

Instructions

FLMD quick guide

FLMD example template

Download FLMD template

CSV Data Dictionary

License

## Overview

File-level metadata provides granular information at the data file level to enable comparison of data files within a data set and the ability to search for and locate files across the data collection. The recommended file-level metadata (FLMD) schema will describe the contents, scope, and structure of the data file within the ESS-DIVE repository. This metadata is fully consistent with and augments the metadata collected to describe each data set.

### Getting started

Instructions for how to use this reporting format:

- File-level metadata reporting format instructions

Other documents:

Copy link

ON THIS PAGE

Getting started

Updates in v1.0.0

How to contribute

Copyright information

Funding and acknowledge...

Recommended citation

Related reference

References

# Questions?

***If you've used the FLMD reporting format, have you had any notable pain points? / If you haven't used the FLMD reporting format, were any fields unclear?***

***Are there any tools you feel could be helpful in using the FLMD reporting format? / Do you use any tools already?***

# CSV Reporting Format

# The CSV Reporting Format

## What is the CSV reporting format?

- The CSV file is a non-proprietary format for tabular data
- Archives tabular data in its simplest form
- Defines structure and some content

## Why use the CSV reporting format?

- Specifies common format for elements within your CSV files (e.g., missing values) which make CSVs easier to read
- Reduces inconsistencies (e.g., 2021-04-26 vs. 4/26/2021)

Terri Velliquette
Oak Ridge National Laboratory

# File Structure

## Character Set

Use the **standard US-ASCII character** set without extensions **or use UTF-8** (which includes the ASCII character set).

*Using either of these character encodings will increase machine readability and interoperability.*

| Ascii | Char | Ascii | Char | Ascii | Char | Ascii | Char |
|---|---|---|---|---|---|---|---|
| 0 | Null | 32 | Space | 64 | @ | 96 | ` |
| 1 | Start of heading | 33 | ! | 65 | A | 97 | a |
| 2 | Start of text | 34 | " | 66 | B | 98 | b |
| 3 | End of text | 35 | # | 67 | C | 99 | c |
| 4 | End of transmit | 36 | $ | 68 | D | 100 | d |
| 5 | Enquiry | 37 | % | 69 | E | 101 | e |
| 6 | Acknowledge | 38 | & | 70 | F | 102 | f |
| 7 | Audible bell | 39 | ' | 71 | G | 103 | g |
| 8 | Backspace | 40 | ( | 72 | H | 104 | h |
| 9 | Horizontal tab | 41 | ) | 73 | I | 105 | i |
| 10 | Line feed | 42 | * | 74 | J | 106 | j |
| 11 | Vertical tab | 43 | + | 75 | K | 107 | k |
| 12 | Form feed | 44 | , | 76 | L | 108 | l |
| 13 | Carriage return | 45 | - | 77 | M | 109 | m |
| 14 | Shift in | 46 | . | 78 | N | 110 | n |
| 15 | Shift out | 47 | / | 79 | O | 111 | o |
| 16 | Data link escape | 48 | 0 | 80 | P | 112 | p |
| 17 | Device control 1 | 49 | 1 | 81 | Q | 113 | q |
| 18 | Device control 2 | 50 | 2 | 82 | R | 114 | r |
| 19 | Device control 3 | 51 | 3 | 83 | S | 115 | s |
| 20 | Device control 4 | 52 | 4 | 84 | T | 116 | t |
| 21 | Neg. acknowledge | 53 | 5 | 85 | U | 117 | u |
| 22 | Synchronous idle | 54 | 6 | 86 | V | 118 | v |
| 23 | End trans. block | 55 | 7 | 87 | W | 119 | w |
| 24 | Cancel | 56 | 8 | 88 | X | 120 | x |
| 25 | End of medium | 57 | 9 | 89 | Y | 121 | y |
| 26 | Substitution | 58 | : | 90 | Z | 122 | z |
| 27 | Escape | 59 | ; | 91 | [ | 123 | { |
| 28 | File separator | 60 | < | 92 | \ | 124 | \| |
| 29 | Group separator | 61 | = | 93 | ] | 125 | } |
| 30 | Record separator | 62 | > | 94 | ^ | 126 | ~ |
| 31 | Unit separator | 63 | ? | 95 | _ | 127 | Forward del. |

## Delimiter

Save files in **CSV format.**

*This requirement is necessary for machine readability as unprotected commas will disrupt the interpretation of columns and rows.*

# File Structure

## Data Matrix

The contents of the data portion of the file must be in a **logical and readable matrix format**. There can be **no empty rows**. There must be the **same number of columns across all of its rows**.

| | | |
|---|---|---|
| row 1 column 1 | row 1 column 2 | row 1 column 3 |
| row 2 column 1 | row 2 column 2 | row 2 column 3 |
| row 3 column 1 | row 3 column 2 | row 2 column 3 |

# File Structure

## Column or row name orientation

The orientation of the Column/Row Names in the Data Matrix could be presented:

1. Horizontally with Names at the top of each column

   OR

2. Vertically with Names at the start of each row.

Horizontal Orientation

| area | plot_type | Latitude | Longitude | year | CH4_flux | C_CO2eq |
|---|---|---|---|---|---|---|
| Site 6 | CLC1 | 71.29573 | -156.66473 | 2010-07-07 | 91.8 | 1.355 |
| Site 6 | CLC2 | 71.29571 | -156.66469 | 2010-07-07 | 54.3 | 1.178 |

Vertical Orientation

| | | | | |
|---|---|---|---|---|
| area | Site 6 | Site 6 | Site 6 | Site 6 |
| plot_type | CLC1 | CLC2 | CLC3 | CLC5 |
| Latitude | 71.29573 | 71.29571 | 71.2957 | 71.28615 |
| Longitude | -156.66473 | -156.66469 | -156.66467 | -156.59787 |
| year | 2010-07-07 | 2010-07-07 | 2010-07-07 | 2010-07-07 |
| CH4_flux | 91.8 | 54.3 | 63.9 | -9999 |
| C_CO2eq | 1.355 | 1.178 | 0.708 | -9999 |
| thaw_depth | 35 | 38 | 35 | -9999 |

# Naming structure

## File Name

**Unique** and **descriptive** file names about file contents

**data1.csv, data from burned.csv, _plots.csv**

**Burned_plot_veg_2016.csv, SoilPoreWaterHillslope2019.csv**

# Naming structure

## Column/row names

Column or row names should be concise and clear.

Use only letters, numbers, hyphens, and underscores
“ ”
–

Do not start with numbers

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| area | plot_type | Latitude | Longitude | year | CH4_flux |
| Site 6 | CLC1 | 71.29573 | -156.66473 | 2010-07-07 | 91.8 |
| Site 6 | CLC2 | 71.29571 | -156.66469 | 2010-07-07 | 54.3 |
| Site 6 | CLC3 | 71.2957 | -156.66467 | 2010-07-07 | 63.9 |

# Naming structure

## Units

Do not include units in data fields.

Can provide units below the column name as a next row /adjacent to the row name as next columns*
OR only in ***CSV Data Dictionary***

Include “N/A” when units are not applicable

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | area | plot_type | Latitude | Longitude | year | CH4_flux |
| 2 | N/A | N/A | Decimal degrees | Decimal degrees | yyyy-mm-dd | mgC-CH4 m2/day |
| 3 | Site 6 | CLC1 | 71.29573 | -156.66473 | 2010-07-07 | 91.8 |
| 4 | Site 6 | CLC2 | 71.29571 | -156.66469 | 2010-07-07 | 54.3 |
| 5 | Site 6 | CLC3 | 71.2957 | -156.66467 | 2010-07-07 | 63.9 |

**** If providing units directly in data file, use header_rows variable in FLMD to note the additional row***

# **Field Structure:** Temporal Data/Range & Spatial Data

- Temporal Data
  - Date reported in ISO 8601 standard: **YYYY-MM-DD**
  - Report to known precision (e.g., YYYY-MM, YYYY)
  - Time reported in UTC: hh:mm:ss
  - Report to known precision (e.g., hh:mm, hh)
  - If date and time are split between two fields, name fields “date” and “time”
- Temporal Data Range
  - Range time stamped data to be reported as paired columns or rows for start and stop times
  - “dateTime_start" and "dateTime_end" OR "time_start" and "time_end"
- Spatial Data
  - Geographic coordinates to be reported in WGS84 decimal format
  - Provide latitude and longitude as separate variables

# **Field Structure:** Consistent Values & Missing Value Codes

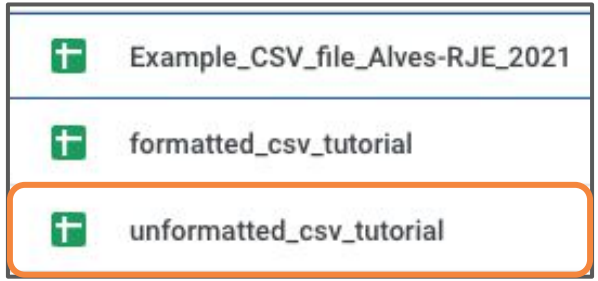
- Consistent Values
  - All data within the Column or Row must use the same units of measurement
  - Do not mix text and numeric data within same Column or Row

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | area | plot_type | Latitude | Longitude | year | CH4_flux | C_CO2eq | thaw_depth | water_table_depth |
| 2 | N/A | N/A | Decimal degrees | Decimal degrees | yyyy-mm-dd | mgC-CH4 m2/day | gC-CO2 m2/day | cm | cm |
| 3 | Site 6 | CLC1 | 71.29573 | -156.66473 | 2010-07-07 | 91.8 | 1.355 | 35 | standing water |
| 4 | Site 6 | CLC2 | 71.29571 | -156.66469 | 2010-07-07 | 54.3 | 1.178 | > probe depth | 1 |
| 5 | Site 6 | CLC3 | 71.2957 | -156.66467 | 2010-07-07 | 63.9 | 0.708 | 35 | 0.5 |

- Missing Value Codes
  - Cells with missing values should be represented with Missing Value Codes
    - Numeric Data = “-9999”
    - Character Data (text) = “N/A”

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | area | plot_type | Latitude | Longitude | year | CH4_flux | C_CO2eq | thaw_depth | water_table_depth | notes |
| 2 | N/A | N/A | Decimal degrees | Decimal degrees | yyyy-mm-dd | mgC-CH4 m2/day | gC-CO2 m2/day | cm | cm | N/A |
| 3 | Site 6 | CLC1 | 71.29573 | -156.66473 | 2010-07-07 | 91.8 | 1.355 | 35 | 0 | rain, ponding |
| 4 | Site 6 | CLC2 | 71.29571 | -156.66469 | 2010-07-07 | 54.3 | 1.178 | 38 | 1 | N/A |
| 5 | Site 6 | CLC3 | 71.2957 | -156.66467 | 2010-07-07 | 63.9 | 0.708 | 35 | 0.5 | N/A |
| 6 | Site 6 | CLC5 | 71.28615 | -156.59787 | 2010-07-07 | -9999 | -9999 | -9999 | -9999 | lost data |

# Practice: CSV Reporting Format


| | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| 1 | **Sample Name** | **Date** | **Time Collected** | **Water Temperature Celsius** | **Dissolved Oxygen mg/L** | **Electrical Conductivity µS** | **Notes** |
| 2 | Sample_Site_1 | 2022-01-12 | 1:05 PM | 22 C | 10.05 | 46.3 uS | |
| 3 | Sample_Site_2 | 01/12/2022 | 13:50 | 20.7 | | 45.5 | |
| 4 | Sample_Site_3 | 01-12-22 | 14:22 | 19.7 | Water level too low to test DO | 54.5 uS | |
| 5 | Sample_Site_4 | 2022-01-12 | 14:56 | | 10.56 | 45 µS | |
| 6 | Sample_Site_5 | 2022-01-12 | 3:12 PM | 21.9 | 9.89 | | |
| 7 | Sample_Site_6 | 01/12/2022 | 16:04 | | 11.01 | 45.2 | Thermome |

# Practice: CSV Reporting Format

## Unformatted CSV Tutorial file (unformatted_csv_tutorial)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Sample Name** | **Date** | **Time Collected** | **Water Temperature Celsius** | **Dissolved Oxygen mg/L** | **Electrical Conductivity µS** |
| 2 | Site 1 Sample | 2022-01-12 | 1:05 PM | 22 C | 10.05 | 46.3 uS |
| 3 | Site 2 Sample | 01/12/2022 | 13:50 | 20.7 | | 45.5 |
| 4 | Site 3 Sample | 01-12-22 | 14:22 | 19.7 | Water level too low to test DO | 54.5 uS |
| 5 | Site 4 Sample | 2022-01-12 | 14:56 | | 10.56 | 45 µS |
| 6 | Site 5 Sample | 2022-01-12 | 3:12 PM | 21.9 | 9.89 | |
| 7 | Site 6 Sample | 01/12/2022 | 16:04 | | 11.01 | 45.2 |

# Practice: CSV Reporting Format

- **Character Set:** Use US-ASCII (includes all upper- and lowercase characters, digits, and common punctuation used in the English language
- **Column/Row Orientation:** data can be presented either vertically or horizontally
- **File, Column, and Row Names:** unique and detailed names; only use letters, numbers, hyphens, and underscores; no spaces
- **Units:** include as next column or row; include N/A when units are not applicable
- **Consistent Values:** include same unit of measurement within column or row; do not mix numeric and text data
- **Missing Value Codes:** N/A for character data (text) and -9999 for numeric
- **Temporal Data:** Format in YYYY-MM-DD hh:mm:ss, to known precision
- **Spatial Data:** Format in WGS84; Provide latitude and longitude in separate columns/rows

***Feel free to review the [CSV Quick Guide]() as you are working through the CSV*

# CSV Reporting Format

## Formatted CSV Data File

| Sample_Name | Date | Water_Temperature | Dissolved_Oxygen | Electrical_Conductivity | Notes |
|---|---|---|---|---|---|
| site_1 | 2022-01-12 | 22 | 10.05 | 46.3 | N/A |
| site_2 | 2022-01-12 | 20.7 | -9999 | 45.5 | N/A |
| site_3 | 2022-01-12 | 19.7 | -9999 | 54.5 | Water level too low to test |
| site_4 | 2022-01-12 | -9999 | 10.56 | 45 | N/A |
| site_5 | 2022-01-12 | 21.9 | 9.89 | -9999 | N/A |
| site_6 | 2022-01-12 | -9999 | 11.01 | 45.2 | Thermometer ran out of power |

***Variable information moved to data dictionary***

# Practice: CSV Reporting Format

- **Character Set:** Use US-ASCII (includes all upper- and lowercase characters, digits, and common punctuation used in the English language
- **Column/Row Orientation:** data can be presented either vertically or horizontally
- **File, Column, and Row Names:** unique and detailed names; only use letters, numbers, hyphens, and underscores; no spaces
- **Units:** include as next column or row; include N/A when units are not applicable
- **Consistent Values:** include same unit of measurement within column or row; do not mix numeric and text data
- **Missing Value Codes:** N/A for character data (text) and -9999 for numeric, unless otherwise defined in the data dictionary
- **Temporal Data:** Format in YYYY-MM-DD hh:mm:ss, to known precision
- **Spatial Data:** Format in WGS84; Provide latitude and longitude in separate columns/rows

# CSV reporting format documentation

Guide to Using ESS-DIVE

Main Website | Search Data | Submit Data | Contact Us

Search...

Welcome

Frequently Asked Questions

SUBMIT DATA

Get Started

Register to Submit Data

Dataset Requirements

Describe & Format Datasets

Submit Data with Online Form

Submit Data with API

Collaborate on Datasets

## CSV File Structure

| CSV File Structure | Status: **READY TO USE** |
|---|---|
| Purpose | Guidance for formatting CSV files submitted to ESS-DIVE |
| Who should use | Anyone submitting tabular data stored in CSV files to ESS-DIVE |
| Documentation | GitHub or GitBook |
| Authors | Terri Velliquette; Ranjeet Devarakonda; Jessica |

Copy link

ON THIS PAGE



CSV file structure reporting format

Overview

Instructions

Documents

## Overview

Tabular data in the form of rows and columns should be archived in its simplest form. Submit these data following the ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure. The CSV reporting format is more likely accessible by future systems over a proprietary format and is preferred because this format is easier to exchange between different programs increasing the interoperability of the data file. Defining the reporting format and structure of the CSV file and some field content increases the machine-readability of the data file for extracting, compiling, and comparing the data across files and systems.

Copy link

ON THIS PAGE



## Getting started

# Questions?

***Pain points? Unclear guidelines?***

# Samples Reporting Format

# ESS-DIVE Sample ID and Metadata Guide and Template

1) *ESS-DIVE documentation for samples*
   *https://ess-dive.gitbook.io/sample-id-and-metadata/*
2) *Instructions* - ***download sample metadata template***
3) *Access metadata guide*
4) *Citation / References* →


Damerow *et al.* 2021.
http://doi.org/10.5334/dsj-2021-011

# ESIP Guide on Publishing Sample-Based Research

Earth Science Information Partners (ESIP)

Physical Samples Curation Cluster

Paper on guide and community/technical needs

A Scientific Author Guide for Publishing Open Research Using Physical Samples

This checklist guide will help authors of scientific papers make their Sample-based studies Open, and Findable, Accessible, Interoperable, and Reusable (FAIR) to advance Sample-based science in the future.


**Author Guide:** https://doi.org/10.6084/m9.figshare.24669057.v1 **Flyer:** https://doi.org/10.6084/m9.figshare.24291148.v2

# Step 1. Describe Samples with Rich Metadata

ESS-DIVE

## Sample Collections Details

- Collector/Chief Scientist*
- Collection Date*
- Collection Time
- Collection Method Description*
- Sample Processing (MIxS)
- Field Program or Project Name*

## Sample Access

- Release Date*
- Current Archive
- Current Archive Contact

## Location

- Location Description
- Latitude*
- Longitude*
- Geolocation Instrument
- Elevation (start, end)
- Elevation Unit
- Country*
- Minimum/Maximum Depth in Meters (DwC)
- Minimum/Maximum Distance above Surface in Meters (DwC)

## Environmental Context

- Physiographic Feature* (ENVO, MIxS)
- Biome (MIxS)

## Sample Description

- *IGSN-SESAR provides*
- Sample Name*
- Object Type* (BCO)
- Material* (ENVO, PO)
- Classification
- Sample Description
- Purpose
- Size, Size Unit
- Filter Size (MIxS)
- Scientific Name (DwC)
- Sample Remarks

## Related Identifiers

- Parent IGSN
- Collection ID (DwC)
- Event ID (DwC)
- Location ID (DwC)

EESA20-061

# Example ESS-DIVE IGSN Template

| Object Type: | Individual Sample | User Code: | IEWDR | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sample Name** | IGSN | Parent IGSN | Release Date | Material | Field name (informal classification) | Collection method | Collection method description | Comment | Latitude | Longitude | Primary physiogra |
| **CM_001_Water** | 10.58052/IEWDR01LH | 10.58052/IEWDR01KZ | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 46.7322 | -117.1805 | stream |
| **CM_002_Water** | 10.58052/IEWDR01LI | 10.58052/IEWDR01KN | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 38.7819 | -77.3884 | stream |
| **CM_003_Water** | 10.58052/IEWDR01LJ | 10.58052/IEWDR01L4 | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 40.1047 | -77.1803 | stream |
| **CM_004_Water** | 10.58052/IEWDR01LK | 10.58052/IEWDR01KK | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 42.7231 | -73.1978 | stream |
| **CM_005_Water** | 10.58052/IEWDR01LL | 10.58052/IEWDR01KY | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 48.8186 | -122.5806 | stream |
| **CM_006_Water** | 10.58052/IEWDR01LM | 10.58052/IEWDR01L2 | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 38.9746 | -119.8218 | stream |
| **CM_007_Water** | 10.58052/IEWDR01LN | 10.58052/IEWDR01KX | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 39.0073 | -121.5804 | stream |
| **CM_008_Water** | 10.58052/IEWDR01LO | 10.58052/IEWDR01L1 | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 33.9485 | -117.6118 | stream |
| **CM_010_Water** | 10.58052/IEWDR01LP | 10.58052/IEWDR01KQ | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 33.6666 | -79.8467 | stream |
| **CM_011_Water** | 10.58052/IEWDR01LQ | 10.58052/IEWDR01KM | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 42.5055 | -73.5062 | stream |
| **CM_012_Water** | 10.58052/IEWDR01LR | 10.58052/IEWDR01KL | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 42.7638 | -73.3372 | stream |
| **CM_013_Water** | 10.58052/IEWDR01LS | 10.58052/IEWDR01L6 | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 41.3093 | -83.1578 | stream |
| **CM_014_Water** | 10.58052/IEWDR01LT | 10.58052/IEWDR01KO | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 35.3668 | -77.4403 | stream |
| **CM_015_Water** | 10.58052/IEWDR01LU | 10.58052/IEWDR01L3 | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 39.4803 | -106.0468 | stream |
| **CM_016_Water** | 10.58052/IEWDR01LV | 10.58052/IEWDR01KR | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 43.4099 | -83.965 | stream |
| **CM_017_Water** | 10.58052/IEWDR01LW | 10.58052/IEWDR01KT | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 45.5974 | -110.566 | stream |
| **CM_018_Water** | 10.58052/IEWDR01LX | 10.58052/IEWDR01LB | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 41.3166 | -102.126 | stream |
| **CM_020_Water** | 10.58052/IEWDR01LY | 10.58052/IEWDR01LC | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 33.8277 | -79.0449 | stream |
| **CM_021_Water** | 10.58052/IEWDR01LZ | 10.58052/IEWDR01LZ | | Liquid>aqueous | Surface water | grab | Surface water was either (1) pulled into syringe fro | WHONDRS CONUS-Scale Model- | 32.3505 | -111.0953 | stream |

Forbes et al., 2023 doi:10.15485/1923689

# Step 2: Assign and Use Identifiers for Samples

## Unique Identifier

Provides a meaningful, project-specific unique ID to organize your data

**Sample Name:**
RockCr001_2021-05-25

## Persistent Identifiers

Globally unique IDs with permanent link/landing page, associated metadata

**ORCiD:** People
**DOI:** Data, publications

**IGSN:** Samples
IEWFS000U

# When do you need PIDs?

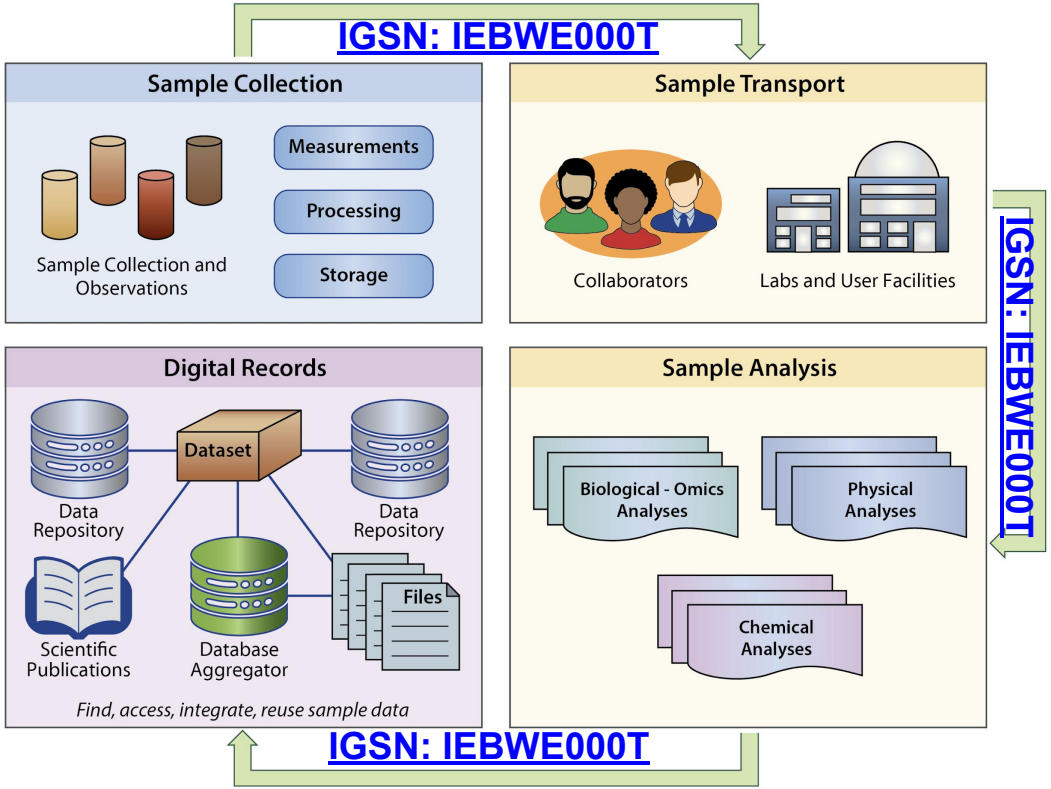
1.) Multiple datasets, journal publication

2.) Collaborators work on same samples

3.) Multiple labs for analyses

4.) Sample-related data in different repositories

5.) Archived, and used for multiple purposes over time



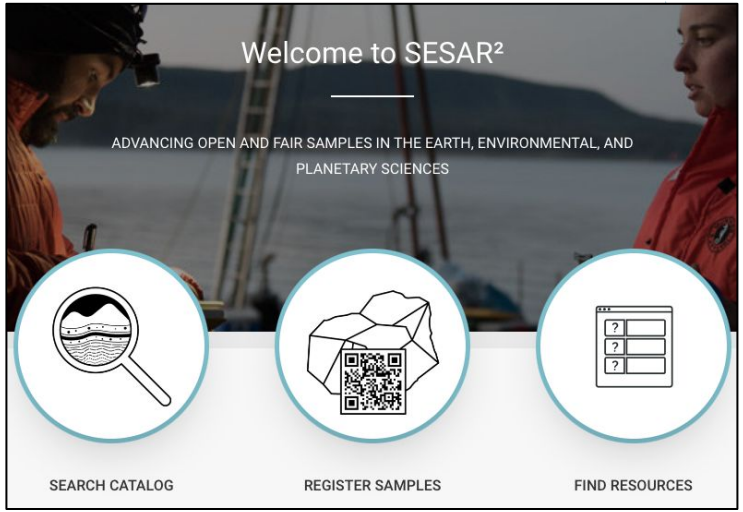

EESA20-037

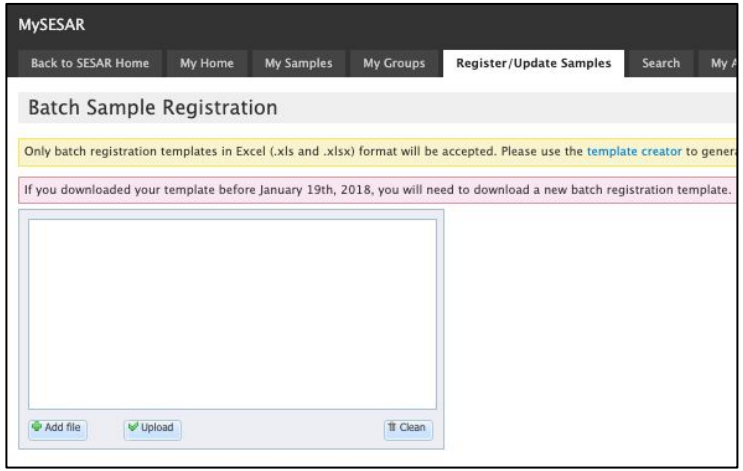
# Step 2. Assign and Use Identifiers for Samples

Register samples for IGSN IDs through SESAR. https://www.geosamples.org/

For subsamples sent to a lab for analysis:

- Provide the laboratory your source material sample PID (IGSN)




nmdc National Microbiome Data Collaborative

EMSL

JGI JOINT GENOME INSTITUTE

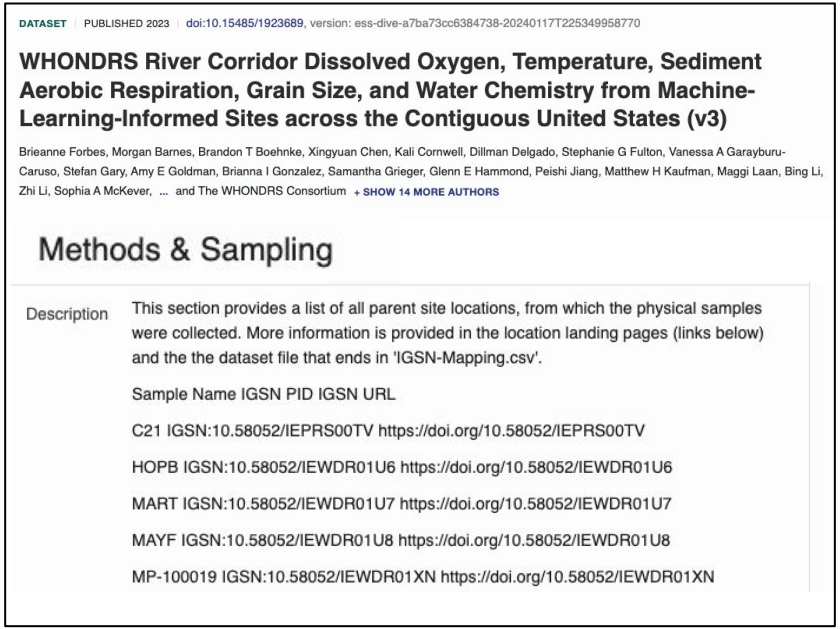
GOLD Genomes Online Database

# Step 3. Publish and Cite Datasets with Sample Identifiers

Publish a dataset that includes your Sample identifiers (ideally PIDs) and associated data

- Include IGSNs your dataset(s) metadata
- Include an IGSN column within all data files containing your sample data

Cite ESS-DIVE dataset(s) in your papers



DATASET | PUBLISHED 2023 | doi:10.15485/1923689, version: ess-dive-a7ba73cc6384738-20240117T225349958770

**WHONDRS River Corridor Dissolved Oxygen, Temperature, Sediment Aerobic Respiration, Grain Size, and Water Chemistry from Machine-Learning-Informed Sites across the Contiguous United States (v3)**

Brieanne Forbes, Morgan Barnes, Brandon T Boehnke, Xingyuan Chen, Kali Cornwell, Dillman Delgado, Stephanie G Fulton, Vanessa A Garayburu-Caruso, Stefan Gary, Amy E Goldman, Brianna I Gonzalez, Samantha Grieger, Glenn E Hammond, Peishi Jiang, Matthew H Kaufman, Maggi Laan, Bing Li, Zhi Li, Sophia A McKever, ... and The WHONDRS Consortium + SHOW 14 MORE AUTHORS

Methods & Sampling

Description This section provides a list of all parent site locations, from which the physical samples were collected. More information is provided in the location landing pages (links below) and the the dataset file that ends in 'IGSN-Mapping.csv'.

Sample Name IGSN PID IGSN URL

C21 IGSN:10.58052/IEPRS00TV https://doi.org/10.58052/IEPRS00TV

HOPB IGSN:10.58052/IEWDR01U6 https://doi.org/10.58052/IEWDR01U6

MART IGSN:10.58052/IEWDR01U7 https://doi.org/10.58052/IEWDR01U7

MAYF IGSN:10.58052/IEWDR01U8 https://doi.org/10.58052/IEWDR01U8

MP-100019 IGSN:10.58052/IEWDR01XN https://doi.org/10.58052/IEWDR01XN

Forbes et al., 2023 doi:10.15485/1923689

# Next Steps for ESS-DIVE Samples

Samples RF does not follow csv RF guidelines
- Need tool to read in Fusion DB

Some minor updates to RF fields not used in SESAR
- Same process as FLMD

Exploring Data Harmonizer tool for validating ESS-DIVE sample metadata

File Settings Validate Help Template EssDive Loaded file ec1_metadata_igsn.csv Next Error

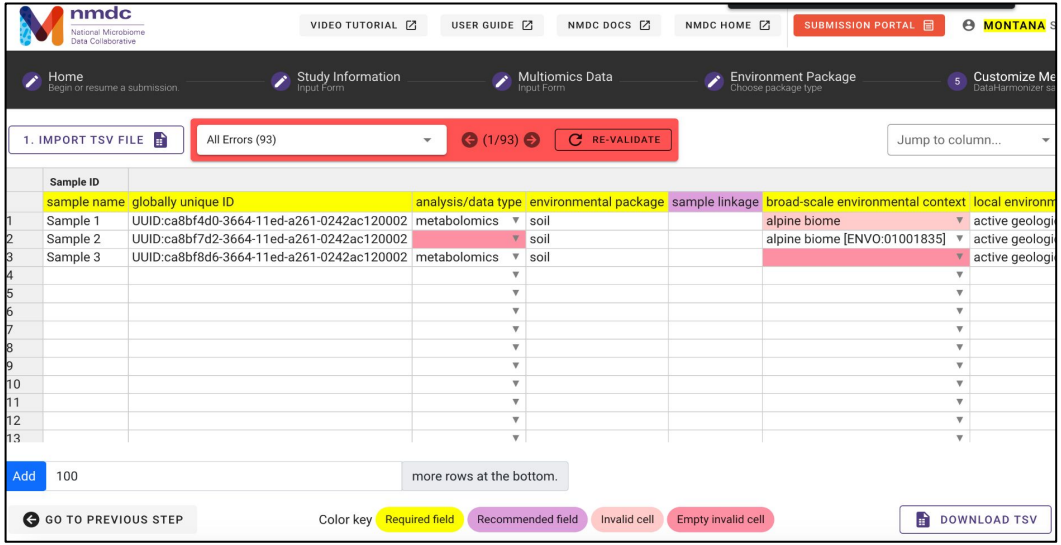
| | Sample IDs and Related Ider | | Sample Collection Details | | | | | | Location | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample Name | nt | Collector/Chief Scientist | Collection date | Collection time | Collection method description | Sample processing | Field program/cruise | Latitude | Longitude | N |
| 1 | EC1_K001_WATER_40ML_FILT | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 2 | EC1_K001_WATER_15ML_FILT | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 3 | EC1_K001_WATER_1L_UNFILT | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 4 | EC1_K001_WATER_125ML_UNFILT | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 5 | EC1_K001_WATER_FILTER | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 6 | EC1_K001_SEDIMENT_JAR | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 7 | EC1_K001_SEDIMENT_BAG | | Donnie Day | 12/14/2021 | | | | | 41.6228 | -83.2362 | |
| 8 | EC1_K001_WETLAND_JAR | | Donnie Day | 12/14/2021 | | | | | 41.62182 | -83.23883 | |
| 9 | EC1_K001_WETLAND_BAG | | Donnie Day | 12/14/2021 | | | | | 41.62182 | -83.23883 | |
| 10 | EC1_K001_WETLAND_RING | | Donnie Day | 12/14/2021 | | | | | 41.62182 | -83.23883 | |
| 11 | EC1_K001_UPLAND_JAR | | Donnie Day | 12/14/2021 | | | | | 41.61511 | -83.22979 | |

# Next Steps for ESS-DIVE Samples

Incorporate ESS-DIVE Samples RF into NMDC Sample Submission

Linking BER data - Related Identifiers

RDA Complex Citations Hackathon for Samples (April 29, 1-3 pm PT)



# Questions?

# Publishing Datasets with Reporting Formats

# File-Level Review and Validation

Datasets using reporting formats go through a second level of validation and review by the Fusion DB ***at the time of publication request.***

Validation is re-run if a dataset files are updated.

| Keywords | CATEGORICAL:NONE |
| --- | --- |
| | **Keyword** |
| | River corridor model |
| | Hyporheic zone |
| | Aerobic respiration |
| | Anaerobic respiration |
| | CRB |
| | Watershed |
| | ESS-DIVE CSV File Formatting Guidelines Reporting Format |
| | ESS-DIVE File Level Metadata Reporting Format |

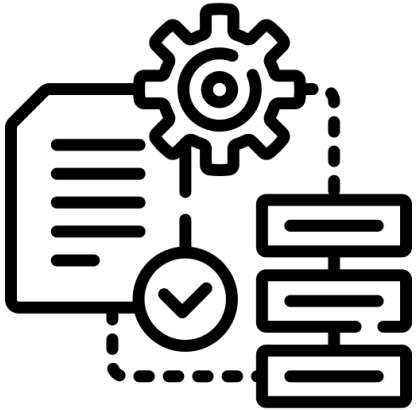
***The Fusion DB uses keywords to identify datasets that have FLMD files***

# File-Level Review and Validation

Fusion DB checks for the inclusion of required fields in FLMD and Data dictionary files, and that CSV files conform to CSV reporting format requirements.*

## Common Errors

- Incorrect naming of required FLMD and DD fields
- Parsing issues in CSVs: variables with spaces or special characters, UTF-8 errors
- Incorrect data orientation defined in FLMD



***Exploring programmatically fixing common errors during review***

# File-Level Review and Validation

Errors are reported to the ESS-DIVE team and sent to data contributor along with any other necessary revision requests

- Errors that are required for the parsing of data files should be fixed before publication
- Files that do parse after publication are available in deep dive

***Considering development of additional methods/external tools for providing feedback before submitting and requesting publication.***

# Resources

Community GitHub - instructions, templates, feedback

Reporting Format Checklist

Past webinars

Portal of datasets using reporting formats for examples

Deep dive API

***Please contact ess-dive-support@lbl.gov with questions or feedback***

# Questions?

# Connect With ESS-DIVE

**To get help:**

ess-dive.lbl.gov

ess-dive-support@lbl.gov

docs.ess-dive.lbl.gov

**To stay updated:**

ess-dive-community@lbl.gov

@essdive

https://bit.ly/essdiveMailingList


**Acknowledgements**

Advisory Groups: ESS-DIVE Archive Partnership Board, ESS Cyberinfrastructure Working Groups

Funding: EESSD Data Management